ISSN No. 0976-5697

Volume 4, No. 10, Sep-Oct 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Predicting Course and Branch Interest among Higher Education Students from Rural and Semi-Urban Area using Data Mining Techniques

Pooja Shrivastava*
Research Scholar,
Barkatullah University, Bhopal
poojarajneeshkarn@gmail.com

Anil Rajput
Professor, Department of Mathematics and Computer Science CSA Govt. PG Nodal College, Sehore (M. P.) India
dranilrajput@hotmail.com

Abstract: - Few years ago, the information flow in education field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that educational institutions face is the explosive growth of educational data and to use this data to extract some useful hidden patterns to improve the quality of education. Data mining is a significant analytical tool for helping organizations to enhance decision making and analyzing new patterns and relationships among a large amount of data. Data analysis plays an important role for any type of decision support irrespective of type of industry. This paper addresses the applications of data mining in educational institution of Madhya Pradesh to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples.

Keywords- Higher education,, Data mining, Knowledge discover, Classification, Association rule, Decision Tree Prediction

I. INTRODUCTION

In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Education, Business, Agriculture and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information.

One of the biggest challenges that higher education faces today is predicting the paths of students admitting in particular courses of affiliated colleges of particular university. Institutions would like to know, which students will enroll in particular course programs and from which region or state they belong, and which students will need assistance in order to graduate. In addition to this challenge, traditional issues such as enrollment management and time-to-degree continue to motivate higher education institutions to search for better solutions. One way to effectively address these student and alumni challenges is through the analysis and presentation of data, or data mining. Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining may, for example, give an institution the information necessary to take action before a student seeks admission or drops out, or to efficiently allocate resources with an accurate estimate of how many students will take a particular course. This paper addresses the capabilities of data mining and its applications in higher education.

Various domestic and foreign researchers have worked a lot in the field of educational data mining which is described in brief as follows:

Brijesh Kumar Baradwaj et al [1,2] described that the hidden patterns, associations and anomalies, which are discovered by some Data mining techniques, can be used to improve the effectiveness, efficiency and the speed of the processes. He used the classification task on student database to predict the students division on the basis of previous database.

Pandey and Pal et al [3,4] conducted study on the student performance based by selecting 60 students from a degree college of Dr. R. M. L. Awadh University, Faizabad, India. By means of association rule they find the interestingness of student in opting class teaching.

K. H. Rashan et. al., [5] discussed that how data mining can help to improve an education system by enabling better understanding of the students. The extra information can help the teachers to manage their classes better and to provide proactive feedback to the students.

Tripti Arjariya *et. al.* [6] described that the DM is directly associated with use of technology for accessing data and to give result as required in a desired way. In Indian contest though computer literacy among the users are very low but its applicability in different sectors of the society is highly demandable day to day. With specific to education sector it has great demand both teaching and learning prospects. The management aspects are highly interference by the Information Communication Technology and DM areas. Higher Education system in India, now a day's totally depended on DM majors. The demand and problem solving abilities within the framework of logical argument and accuracy of result need to explore through research and development procedure.

Varun Kumar et al [7] made an empirical study of the applications of data mining techniques in higher education. According to him, it is crucial to have the right information available at the right time. Therefore, it is important to identify the methods and models, which can extract reliable

and comprehensive knowledge from the higher education students data.

Manpreet Singh Bhullar [8] explained that the current education system does not involve any prediction about fail or pass percentage based on the performance. The system doesn't deal with dropouts. There is no efficient method to caution the student about the deficiency in attendance. It doesn't identify the weak student and inform the teacher. Another common problem in larger colleges and universities, some students may feel lost in the crowd. Whether they're struggling to find help with coursework, or having difficulty choosing (or getting into) the courses they need, many students are daunted by the task of working through the collegiate bureaucracy. Since the proposed model identifies the weak students, the teachers can provide academic help for them. It also helps the teacher to act before a student drops or plan for recourse allocation with confidence gained from knowing how many students are likely to pass or fail.

Mohammed M. Abu Tair *et.al.* [9] given a case study in the educational data mining. It showed how useful data mining can be used in higher education particularly to improve graduate students' performance. Authors used graduate students data collected from the college of Science and Technology in Khanyounis. The data include fifteen years period [1993-2007]. Authors applied data mining techniques to discover knowledge. Particularly they discovered association rules and they sorted the rules using lift metric. Then they used two classification methods which are Rule Induction and Naïve Bayesian classifier to predict the Grade of the graduate student. Also authors clustered the students into groups using K-Means clustering algorithm. Finally, authors used outlier detection to detect all outliers in the data, two outlier methods are used which are Distance-based Approach and Density-Based Approach. Each one of these tasks can be used to improve the performance of graduate student.

Monika Goyal *et. al.* [10] discussed some important issues related to business community and education system along with their solutions. Data analysis plays an important role for any type of decision support irrespective of type of industry. Data warehousing and data mining methods for data analysis are explained in detail.

Sushil Verma et al [11] have discussed the various data mining techniques which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of retainment of faculties in the university. Data mining techniques capabilities provided effective improving tools for student performance. It showed how useful data mining can be in higher education in particularly to predict the final performance of student.

Yujie Zheng et al [12] given in his research paper that data mining in recent years with the database and artificial intelligence developed a new technology. As one important function of data mining is clustering analysis which is a separate tool to discover data sources distribution of information,

Bhise R.B. et. al.[13] made a use of data mining process in a student's database using K-means clustering algorithm to predict students result. We hope that the information generated after the implementation of data mining technique may be helpful for a instructor as well as for students.

Muslihah Wook *et. al.* [14] explained that data mining technology gains its importance and accepted as the most promising technology for KMS; particularly in the context of IHLs. Data mining technology able to create valuable knowledge to be used in helping decision making to eliminate repeating previous mistakes of analyzing students' data. Thus, the integration of data mining technology and KMS could also help to enhance accountability, transparency and the smooth running of IHL.

II. DATA MINING DEFINITION AND TECHNIQUES

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The sequences of steps identified in extracting knowledge from data are: shown in Figure 1.

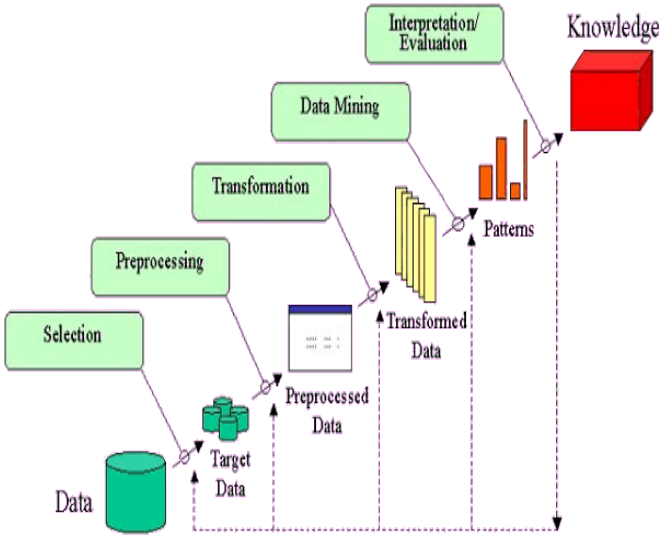

Figure 1. General steps involved in knowledge discovery

The various techniques used in Data Mining are:

A. Decision tree:

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Classification trees are used for the kind of Data Mining problem which are concerned with prediction. Using examples of cases it is possible to construct a model that is able to predict the class of new examples using the attributes of those examples. An experiment has been set up to test the performance of two pruning methods which are used in classification tree modeling. The pruning methods will be applied on different types of data sets.

It is a machine learning technique that allows us to estimate a quantitative target variable (for example, profit, loss or loan amount) or classify observation into one category of a categorical target variable (for example, good/bad credit customer; churn or do not churn) by

repeatedly dividing observations into mutually exclusive groups. The algorithm commonly used to construct decision tree is known as recursive partitioning and the common algorithms are CHAID (Chi-square Automatic Interaction Detection), CART (Classification & Regression Tree) and C5.0. This paper will focus on using CART in building the decision tree. Decision trees represent a supervised approach to classification. Weka uses the J48 algorithm, which is Weka's implementation of C4.5 Decision tree algorithm. J48 is actually a slight improved to and the latest version of C4.5. It was the last public version of this family of algorithms before the commercial implementation C5.0 was released.

A decision tree is a classifier expressed as a recursive partition of the in- stance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Figure 2 describes a decision tree that reasons whether or not a potential customer will respond to a direct mailing. Internal nodes are represented as circles, whereas leaves are denoted as tri- angles. Note that this decision tree incorporates both nominal and numeric at- tributes. Given this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree), and understand the behavioral characteristics of the entire potential customers population regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values.

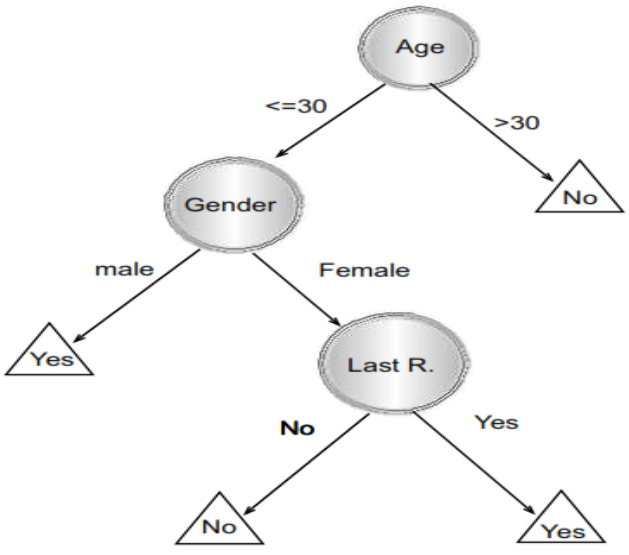

Figure 2 Decision Tree Presenting Response to Direct Mailing

In case of numeric attributes, decision trees can be geometrically interpreted as a collection of hyper planes, each orthogonal to one of the axes. Naturally, decision-makers prefer less complex decision trees, since they may be considered more comprehensible.

The tree complexity has a crucial effect on its accuracy. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used. Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value.

B. Association Rule Generation:

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Based on the concept of strong rules, Agrawal [15] introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onion,potatoes}=>{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.

In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

a. Definition:

Following the original definition by Agrawal [15] the problem of association rule mining is defined as:

Let **I = {i1, i2, ..., in}** be a set of *n* binary attributes called *items*. Let **D = {t1, t2, ..., tn}** be a set of transactions called the *database*. Each transaction in **D** has a unique transaction ID and contains a subset of the items in **I**. A *rule* is defined as an implication of the form **X_Y** where **X, Y** ⊆ **I** and _ ∩ _ = ∅**.** The sets of items (for short *itemsets*) **X** and **Y** are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

C. Classification and Prediction:

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision

trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction.

IF-THEN rules are specified as **IF condition THEN Conclusion**

e.g. IF age=youth and student=yes then buys_computer=yes

D. Clustering Analysis:

Unlike classification and predication, which analyze class labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Application of clustering in education can help institutes group individual student into classes of similar behavior. Partition the students into clusters, so that students within a cluster (e.g. Average) are similar to each other while dissimilar to students in other clusters (e.g. Intelligent, Weak).

III. EXPERIMENTAL RESULTS

Higher Education System of Madhya Pradesh has a great amount of data which can be analyzed and extracted for the data mining system. Faculties and departments have also important detail data regarding courses and modules which will be collected in document form (excel sheets).

In this section, association rules and decision tree data mining techniques and cluster analysis have been used for discovering hidden information from database. For the same purpose, database of Alpha College, Shyampur which is rural area, and Sehore which is semi-uraban area from Madhya Pradesh state have been taken to apply data mining techniques and to calculate results.

From database, first attributes are required which is selected from these excel sheets. These attributes are as follows:

a. Course
b. Branch
c. Gender
d. Category
e. Class
f. Income
g. Date of Admission
h. Minority/Non-Minority

Weka as data mining tool has been used for experimental results. First we open the GUI of Weka which looks like as follows:

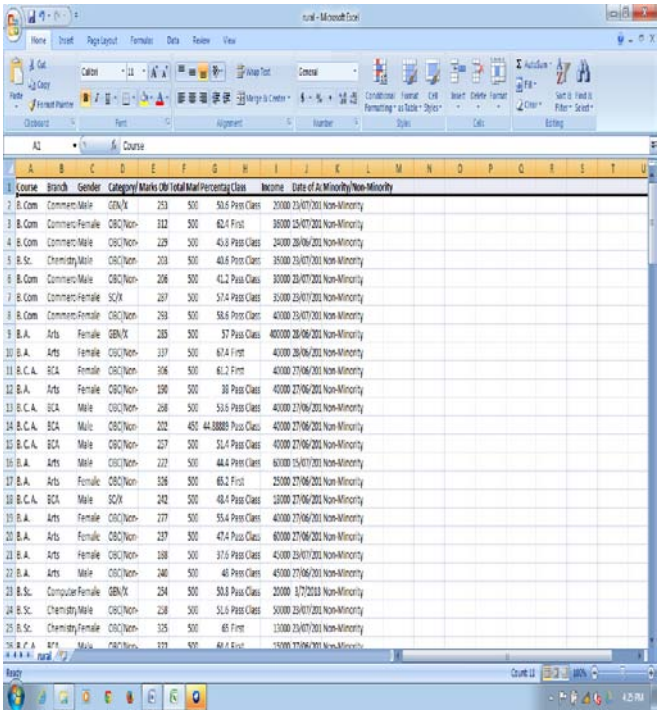

Figure 3 Dataset sample

Then we open Explorer application to preprocess the dataset. We apply the unsupervised filter to convert the attribute values form numeric to nominal as the processing is done on nominal in decision tree.

A. Results from Rural Area:

Data from all excel sheets are collected in one excel sheet and excel sheet is saved in "rural.csv" file for processing in WEKA. In this dataset, we are having total 134 records. Out of this 34% of data is used for training and 66% data is used as testing purpose. Ten fold cross-validations are done on data.

a. Decision Tree when attribute "Course" is selected:

From this decision tree we can easily identify number of students admitted for particular course. From this data we can recognize the most selected and less selected courses by the students.

Number of Leaves : 6
Size of the tree : 7
Time taken to build model : 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	134	100 %
Incorrectly Classified Instances	0	0 %
Total Number of Instances	134	

Confusion matrix is calculated which shows the classification courses i. e. whether they are correctly classified or misclassified.

=== Confusion Matrix ===

a	b	c	d	<-- classified as
20	0	0	0	| a = B. Com
0	18	0	0	| b = B. Sc.
0	0	65	0	|c = B. A.
0	0	0	31	| d = B. C. A.

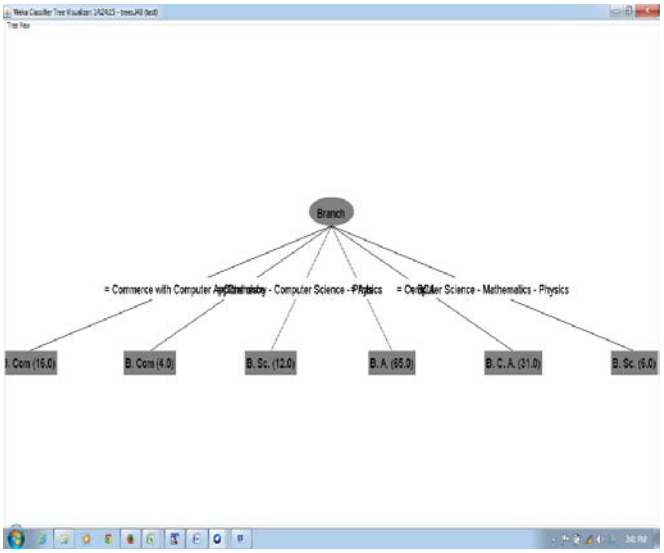

Figure 4 Decision Tree Visualization When Attribute "Course" is Selected

b. Decision Tree when attribute "Branch" is selected:

By doing this analysis we can easily identify the most popular branch among students.

Number of Leaves : 10
Size of the tree : 12
Time taken to build model : 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 123 91.791 %
Incorrectly Classified Instances 11 8.209 %

Confusion matrix is calculated which shows the classification branches i. e. whether they are correctly classified or misclassified.

=== Confusion Matrix ===

a	b	c	d	e	f	<--classified as
16	0	0	0	0	0	a = Commerce with Computer Application
4	0	0	0	0	0	|b = Commerce
0	0	11	0	0	1	|c = Chemistry -Computer Science - Physics
0	0	0	65	0	0	| d = Arts
0	0	0	0	31	0	| e = BCA
0	0	6	0	0	0	| f =Computer Science-Mathematics -Physics

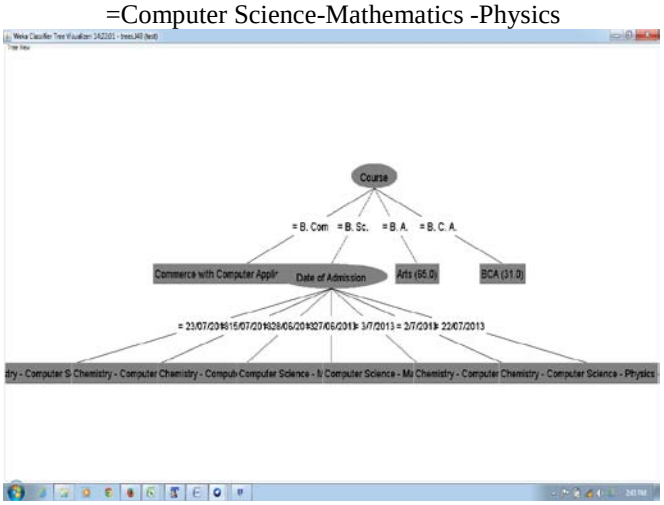

Figure 5. Decision Tree Visualization When Attribute Branch is selected

c. Cluster Analysis when "Date of Admission" Attribute is selected:

When date of admission attribute is selected for cluster formation using clustering algorithm "Cluster using expectation maximization", so we can identify number of students admitted on that particular date.

Class attribute : Date of Admission
Classes to Clusters:

0	1	2	3	4	<--assigned to cluster
15	14	16	0	35	| 23/07/2013
0	0	1	0	1	| 15/07/2013
0	0	3	0	11	| 28/06/2013
2	17	0	0	9	| 27/06/2013
1	0	0	0	0	| 3/7/2013
0	0	0	0	2	| 2/7/2013
0	0	0	1	6	| 22/07/2013

Cluster 0 <-- 3/7/2013
Cluster 1 <-- 27/06/2013
Cluster 2 <-- 28/06/2013
Cluster 3 <-- 22/07/2013
Cluster 4 <-- 23/07/2013

Incorrectly clustered instances: 77.0 57.4627 %

d. Cluster Analysis when "Course" Attribute is selected:

When course attribute is selected for cluster formation using clustering algorithm "Cluster using expectation maximization", it gives how many clusters can be formed and how four courses are assigned to different clusters.

Clustered Instances
0 55 (41%)
1 71 (53%)
2 2 (1%)
3 6 (4%)

Class attribute: Course
Classes to Clusters:

0	1	2	3	<-- assigned to cluster
1	17	2	0	| B. Com
15	2	0	1	| B. Sc.
10	51	0	4	| B. A.
29	1	0	1	| B. C. A.

Cluster 0 <-- B. C. A.
Cluster 1 <-- B. A.
Cluster 2 <-- B. Com
Cluster 3 <-- B. Sc.
Incorrectly clustered instances: 51.0 38.0597 %

e. Cluster Analysis when "Class" Attribute is selected:

From this analysis we can identify how many students are there in Pass class, first class and distinction.

Class attribute: Class
Classes to Clusters:

0	1	<-- assigned to cluster
77	30	| Pass Class
12	5	| Fisrt
9	1	| Distinction

Cluster 0 <-- Pass Class
Cluster 1 <-- Fisrt
Incorrectly clustered instances: 52.0 38.806 %

f. Association Rules

Apriori association rule learner is used for generating rules.

Apriori

Minimum support : 0.45(60instances)
Minimum metric <confidence> : 0.9
Number of cycles performed : 11

a) Generated sets of large itemsets:

Size of set of large itemsets L(1): 8
Size of set of large itemsets L(2): 13
Size of set of large itemsets L(3): 7
Size of set of large itemsets L(4): 1

b) Best rules found:

(a) Branch=Arts 65 ==> Course=B. A. 65 conf:(1)
(b) Course=B. A. 65 ==> Branch=Arts 65 conf:()

Above both rules validate each other that if branch is Arts, then course will be B.A. and it is true for all 65 instances of occurrences of branch Arts.

(c) Course=B. A. 65 ==> Minority/Non-Minority=Non-Minority 65 conf:(1)
(d) Course=B. A. 65 ==> Minority/Non-Minority=Non-Minority 65 conf:(1)
(e) Branch=Arts 65 ==> Minority/Non-Minority=Non-Minority 65 conf:(1)
(f) Branch=Arts Minority/Non-Minority=Non-Minority 65 ==> Course=B. A. 65 conf:(1)
(g) Course=B. A. Minority/Non-Minority=Non-Minority 65 ==> Branch=Arts 65 conf:(1)
(h) Course=B. A. Branch=Arts 65 ==> Minority/Non-Minority=Non-Minority 65 conf:(1)
(i) Branch=Arts 65 ==> Course=B. A. Minority/Non-Minority=Non-Minority 65 conf:(1)
(j) Course=B. A. 65 ==> Branch=Arts Minority/Non-Minority=Non-Minority 65 conf:(1)

Above all rules give the association between course B.A., branch Arts and Minority/Non-Minority which is in this case Non-Minority is true for all 65 cases so that confidence is 1.

B. Results from Semi-urban Area

Data from all excel sheets are collected in one excel sheet and excel sheet is saved in "semiurban.csv" file for processing in WEKA. In this dataset we are having total 1181 records. Out of this 34% of data is use for training and 66% data is used as testing purpose. Ten fold cross-validations are done on data.

a. Decision Tree when attribute "Course" is selected:

From this decision tree we can easily identify number of students admitted for particular course. From this data we can recognize the most selected and less selected courses by the students.

Number of Leaves : 26
Size of the tree : 27
Time taken to build model : 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	1039	87.9763 %
Incorrectly Classified Instances	142	12.0237 %
Total Number of Instances	1181	

Confusion matrix is calculated which shows the classification courses i. e. whether they are correctly classified or misclassified.

=== Confusion Matrix ==

a	b	c	d	e	f	g	h	<-- classified as
314	0	0	0	0	0	0	0	| a = B. A.
0	0	0	18	0	0	0	0	| b = B. C. A.
0	0	0	124	0	0	0	0	| c = B. Com
0	0	0	161	0	0	0	0	| d = B. Sc.
0	0	0	0	360	0	0	0	| e = M. A.
0	0	0	0	0	40	0	0	| f = M. Com.
0	0	0	0	0	0	134	0	| g = M. Sc.
0	0	0	0	0	0	0	30	| h = Special Courses

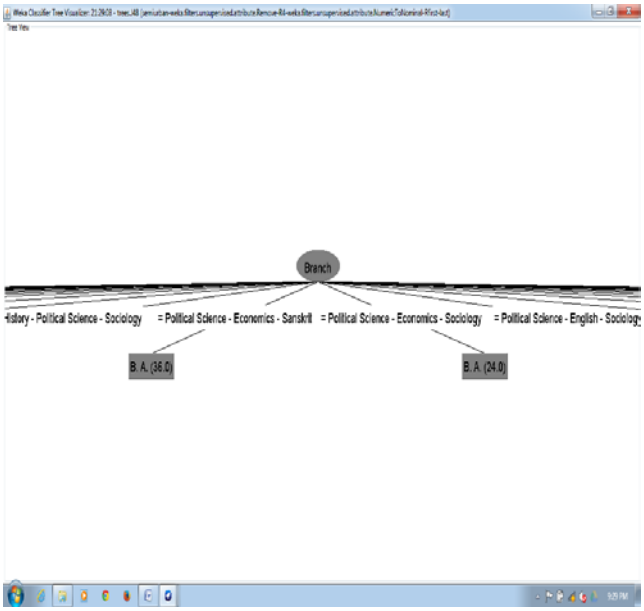

Figure 6 Decision Tree when Course Attribute is selected.

b. Decision Tree when attribute "Branch" is selected:

By doing this analysis we can easily identify the most popular branch among students.

Also we can identify number of students admitted for each branch.

Number of Leaves : 863
Size of the tree : 867
Time taken to build model : 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	759	64.2676 %
Incorrectly Classified Instances	422	35.7324 %
Total Number of Instances	1181	

Confusion matrix is calculated which shows the classification branches i. e. whether they are correctly classified or misclassified.

=== Confusion Matrix ===

a b c d e f g h i j k l m n o p q r s t u v w x y z <-- classified as

0 1 0 0 2 1 0 0 0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | a = NA

1 2 1 2 1 7 1 0 1 1 2 0 6 0 1 0 0 0 0 0 0 0 0 0 0 0 0 | b = Economics - Hindi - Political Science

0 1 0 1 0 1 0 0 0 0 1 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = English - Economics - Sociology

0 2 1 2 1 1 0 1 0 3 2 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | d = Geography - Hindi - Political Science

2 2 0 2 0 2 2 1 0 3 2 0 15 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | e = Geography - Hindi - Sanskrit

2 9 1 1 2 6 0 0 0 3 0 1 4 2 0 0 0 0 0 0 0 0 0 0 0 0 0 | f = Geography - Hindi - Sociology

1 1 0 0 3 1 0 1 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | g = Geography - Political Science - Sociology

0 1 0 1 1 0 1 0 3 1 4 2 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | h = Geography - Sanskrit - Sociology

0 2 0 0 2 0 0 3 2 2 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | i = History - English - Sociology

2 2 0 3 3 1 0 1 1 7 0 0 10 2 0 0 0 0 0 0 0 0 0 0 0 0 0 | j = History - Hindi - Political Science

4 6 1 4 1 2 0 2 0 0 0 3 7 1 1 0 0 0 0 0 0 0 0 0 0 0 0 | k = History - Hindi - Sociology

1 1 0 1 1 2 0 2 1 2 2 0 7 2 1 0 0 0 0 0 0 0 0 0 0 0 0 | l = History - Political Science - Sociology

1 3 1 3 4 1 0 3 0 2 3 1 13 1 0 0 0 0 0 0 0 0 0 0 0 0 0 | m = Political Science - Economics - Sanskrit

0 3 1 2 0 3 0 1 0 5 1 1 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | n = Political Science - Economics - Sociology

0 3 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | o = Political Science - English - Sociology

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 303 0 0 0 0 0 0 0 0 0 0 | p = Applicable for BA only

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 329 0 0 0 0 0 0 0 0 0 | q = Sociology

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 31 0 0 0 0 0 0 0 0 0 | r = Hindi

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 40 0 0 0 0 0 0 0 0 | s = Commerce

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 11 0 | t = Biotechnology

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 23 0 | u = Botany

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 22 0 | v = Chemistry

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 23 0 | w = Physics

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 22 0 | x = Zoology

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 25 0 | y = Mathematics

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 30 | z = PG Dipolma in Computer Applications

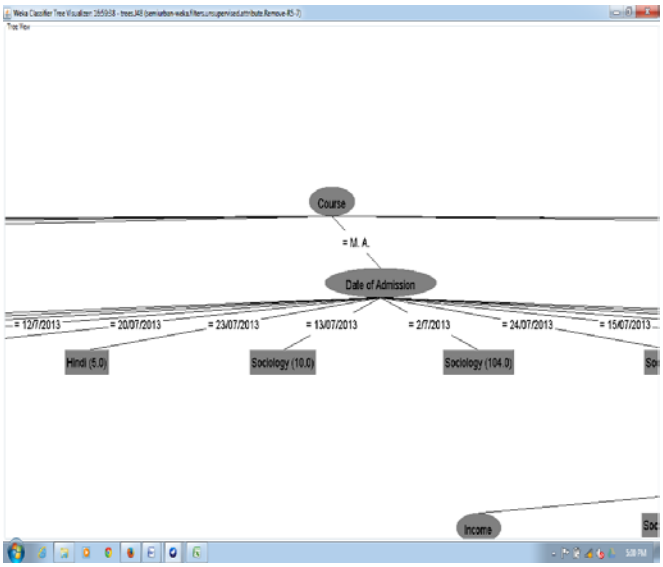

Figure7. Decision Tree when Branch Attribute is selected

c. Cluster Analysis when "Date of Admission" Attribute is selected:

When date of admission attribute is selected for cluster formation using clustering algorithm "Cluster using expectation maximization", so we can identify number of students admitted on that particular date.

Clustered Instances

0 314 (27%)
1 360 (30%)
2 145 (12%)
3 164 (14%)
4 158 (13%)
5 40 (3%)

Class attribute: Date of Admission

Classes to Clusters:

0	1	2	3	4	5<-- assigned to cluster
9	10	0	9	4	1| 1/7/2013
69	91	21	51	31	16| 3/7/2013
1	0	4	0	15	0| 11/7/2013
12	0	3	0	6	0| 12/7/2013
35	67	15	1	21	3| 20/07/2013
56	5	51	14	15	1| 23/07/2013
45	10	5	4	4	0| 13/07/2013

54	104	12	30	17	8| 2/7/2013
16	0	7	6	3	0| 24/07/2013
14	0	6	12	7	3| 15/07/2013
3	51	9	22	24	8| 22/07/2013
0	0	10	10	8	0| 29/06/2013
0	0	2	0	3	0| 28/06/2013
0	22	0	5	0	0| 16/07/2013

Cluster 0 <-- 13/07/2013
Cluster 1 <-- 2/7/2013
Cluster 2 <-- 23/07/2013
Cluster 3 <-- 3/7/2013
Cluster 4 <-- 20/07/2013
Cluster 5 <-- 22/07/2013
Incorrectly clustered instances: 901.0 76.2913 %

d. Cluster Analysis when "Course" Attribute is selected:

When course attribute is selected for cluster formation using clustering algorithm "Cluster using expectation maximization", it gives how many clusters can be formed and how four courses are assigned to different clusters.

Clustered Instances
0 343 (29%)
1 44 (4%)
2 56 (5%)
3 132 (11%)
4 12 (1%)
5 284 (24%)
6 29 (2%)
7 281 (24%)
Class attribute: Course
Classes to Clusters:

0	1	2	3	4	5	6	7	<-- assigned to cluster
64	2	6	70	0	28	9	135	| B. A.
7	2	0	1	0	1	1	6	| B. C. A.
37	2	0	8	0	0	6	71	| B. Com
86	16	1	1	0	0	1	56	| B. Sc.
20	0	37	46	11	246	0	0	| M. A.
19	2	2	3	0	7	2	5	| M. Com.
85	19	10	2	0	1	9	8	| M. Sc.
25	1	0	1	1	1	1	0	| Special Courses

Cluster 0 <-- B. Sc.

Cluster 1 <-- M. Sc.

Cluster 2 <-- M. Com.

Cluster 3 <-- B. Com

Cluster 4 <-- Special Courses

Cluster 5 <-- M. A.

Cluster 6 <-- B. C. A.

Cluster 7 <-- B. A.

Incorrectly clustered instances: 683.0 57.8323 %

e. Cluster Analysis when "Class" Attribute is selected:

From this analysis we can identify how many students are there in Pass class, first class and distinction. So we can concentrate on weak students to improve their performance and help bright students to keep their performance increase.

Clustered Instances
0 791 (67%)
1 390 (33%)
Class attribute: Class
Classes to Clusters:

0	1	<-- assigned to cluster
322	119	| Fisrt
384	235	| Pass Class
85	36	| Distinction

Cluster 0 <-- First
Cluster 1 <-- Pass Class
Incorrectly clustered instances : 624.0 52.8366 %

f. Association Rules:

Apriori association rule learner is used for generating rules.

Apriori

Minimum support: 0.1 (118 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

a) Generated sets of large itemsets:

Size of set of large itemsets L(1): 15
Size of set of large itemsets L(2): 26
Size of set of large itemsets L(3): 7
Size of set of large itemsets L(4): 1

b) Best rules found:

(a) Branch=Sociology 329 ==> Course=M. A. 329 conf:(1)
(b) Branch=Sociology Class= Pass Class 259 ==> Course=M. A. 259 conf:(1)
(c) Branch=Sociology Gender=Male 211 ==> Course=M. A. 211 conf:(1)
(d) Branch=Sociology Gender=Male Class= Pass Class 178 ==> Course=M. A. 178 conf:(1)

The fourth rule is the combination of first three rules i.e. is branch is Sociology, gender is Male and class is Pass Class then course is M.A. which is true for all instances of course M.A. In the 5^{th} rule reverse of 4^{th} is there

(e) Course=B. Sc. 161 ==> Branch=Applicable for BA only 161 conf:(1)
(f) 6. Course=B. Com 124 ==> Branch=Applicable for BA only 124 conf:(1)
(g) 7. Branch=Sociology Gender=Female 118 ==> Course=M. A. 118 conf:(1)
(h) 8. Branch=Sociology Income=36000 118 ==> Course=M. A. 118 conf:(1)

In rule 8^{th} if branch is Sociology and income is 36000 then course is M.A. is true for all 118 instances of M.A.

IV. CONCLUSION

In the present paper, the data mining techniques used are association rules, cluster analysis and decision trees which help to uncover the hidden patterns from the large dataset of higher education students from rural and semi-urban area of Madhya Pradesh. These hidden patterns are useful for both teachers and management. Decision trees are generated depending on various attributes such as course and branch. The information retrieved may help to improve the result of students, the performance of students, students behavior, carefully designing course curriculum to motivate students and to find out the popularity of a particular course among students. Clusters analysis is done on datasets based on attribute date of admission. This gives information about number of students admitted on that particular date. From this we can get statistics about number of students admitted on different dates during admission's time period.

Another cluster is created based on attribute class and course which is grade of student in their previous exam. From this analysis we can figure out number of students with pass class, first class and distinction course wise. So we can concentrate on weak students to improve their performance during academics.

By using Apriori association rule learner rules are generated for each area as well as all combine data which gives if…..then rules with their associated confidence. These rules give information about which course having which branches, which income group selected which branch under which course. When we apply this learner to combine data it gives information about which course is offered in which area with which branches and for how many instances these rules are true is given in the form of confidence associated with each rule.

So we conclude that by using data mining techniques i.e. decision tree, cluster analysis and association rules we can identify hidden information from the given dataset for our benefit.

V. REFERENCES

[1] Bharadwaj, B.K., and Pal, S., "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[2] Brijesh Kumar, B., Sourabh, P.,"Mining Educational Data to Analyze Student's Performance", IJACSA Volume 2, no. 6, 2011.

[3] Pandey, U. K. and Pal, S., "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[4] Pandey, U. K., and Pal, S., "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[5] Rashan, K. H., Anushka, P., "Data Mining Applications in the Education Sector", MSIT, Carnegie Mellon University, retrieved on 28/01/2011.

[6] Tripti, A., Shiv, K., Rakesh, S., Dinesh, V., "Data Mining and It's Approaches towards Higher Education Solutions", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-5, November 2011.

[7] Varun. K., Anupama, C., "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, pp. 80-84, March 2011.

[8] Manpreet Singh, B., Amritpal Kaur, "Use of Data Mining in Education Sector", Proceedings of the World Congress on Engineering and Computer Science 2012,Vol I, WCECS 2012, October 24-26, 2012, San Francisco, USA.

[9] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.

[10] Monika, G. and Rajan, V., "Applications of Data Mining in Higher Education", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814.

[11] Sushil, V., Thakur, R. S. and Shailesh, J.**, "**A Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Computer & Communication Technology (IJCCT) ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 –7449 Vol-3, Iss-3, 2012.

[12] Yujie, Z., "Clustering Methods in Data Mining with its Applications in High Education", International Conference on Education Technology and Computer (ICETC2012), IPCSIT vol.43 (2012) © (2012) IACSIT Press, Singapore.

[13] Bhise, R.B., Thorat, S.S., Supekar, A.K., "Importance of Data Mining in Higher Education System", IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN: 2279-0837, ISBN: 2279-0845. Volume 6, Issue 6 (Jan. - Feb. 2013), PP 18-21.

[14] Muslihah, W., Zawiyah M., Y., and Mohd Zakree Ahmad, N., "Preliminary Overview of Data Mining Technology for Knowledge Management System in Institutions of Higher Learning", World Academy of Science, Engineering and Technology 74 2013.

[15] Agrawal R. and Srikant R., "Fast algorithm for mining association rules", Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[16] B. Manoj and Ojha D.B., "Study of application of data mining techniques in education"International Journal of Research in Science and Technology, vol. 1, no.4, pp: 1-10, march 2012

Short Bio data for the Authors

 Pooja Shrivastava born in 1980 in Sehore, India. She received the Masters degree in Computer Management from Barkatullah University in 2004. Presently, she is working towards the Ph.D. degree in Computer Science from Barkatullah University, Bhopal. Her research interests are in area of application of data mining in improving higher education system.

 Anil Rajput born in 1965 in India. He received the Masters degree in Mathematics and Computer Application from Govt. P.G. College Sehore affiliated to Barkatullah University in 1987. Presently, he is working as professor in mathematics and computer science in C.S.A. Nodal P.G. College, Sehore. His research interests are in area of application of data mining and neural network.